Structure, alternative splicing and chromosomal localization of the cystatin-related epididymal spermatogenic gene

Gail A. CORNWALL[1], Nelson HSIA and H. Gayle SUTTON

Department of Cell Biology and Biochemistry, Texas Tech University Health Sciences Center, 3601 4th Street, Lubbock, TX 79430, U.S.A.

The cystatin superfamily of cysteine protease inhibitors consists of three major families, including the stefins, cystatins and kininogens. However, the recent identification of several genes that possess sequence similarity with the cystatins but have different gene or protein structures indicates that several new cystatin families or subgroups of families might exist. We previously identified the cystatin-related epididymal spermatogenic (*Cres*) gene, which is related to the family 2 cystatins but exhibits highly tissue-specific expression in the reproductive tract. In the studies presented here, an analysis of gene structure as well as chromosomal mapping studies suggest that the *Cres* gene might represent a new subgroup within the family 2 cystatins. Although the *Cres* gene possesses an additional exon encoding 5′ untranslated sequences, its coding exons are similar in size to the three coding exons of the cystatin family 2 genes, and the *Cres* exon/intron splice junctions occur in identical locations as in the cystatin C gene. Furthermore, chromosomal mapping studies show that the *Cres* gene co-segregates with the cystatin C gene on mouse chromosome 2. Similar to the cystatin family 2 proteins, the Cres protein possesses the type A and B disulphide loops that are necessary for cystatin folding. Interestingly, Cres protein also possesses half of a type C disulphide loop. Although probably related to the cystatin genes, the *Cres* gene is distinct in that its promoter contains consensus motifs typical of regulated genes. Finally, reverse transcriptase-mediated PCR studies and the identification of new *Cres* cDNA clones indicate that the *Cres* mRNA is alternatively spliced, resulting in two *Cres* mRNAs that might be involved in the regulation of *Cres* function.

Key words: epididymis, pituitary gland, reproduction, testis.

INTRODUCTION

The cystatins are a superfamily of cysteine protease inhibitors that are conserved throughout evolution. Early studies identified three families of the cystatin superfamily including the stefins, cystatins and kininogens [1]. The stefins (family 1 cystatins) are characterized as intracellular cysteine protease inhibitors of approx. 10 kDa that lack disulphide bonds. The cystatins (family 2 cystatins) are secreted proteins of 12–13 kDa that are found in most body fluids and contain disulphide bonds. The kininogens (family 3 cystatins) are secreted proteins of 68–120 kDa that are primarily found in blood and also have disulphide bonds. The primary structures of several of the cystatin genes and proteins have been described and several consensus regions identified. For example, the family 2 cystatin genes are characterized by their three-exon structure with conserved exon/intron splice sites and a clustering of some of these genes at the same genetic loci [2,3]. Mutation analysis [4] and X-ray crystallography [5,6] have revealed three conserved regions in the cystatin superfamily proteins that are considered to be important for the inhibition of cysteine proteases, including a glycine residue at the N-terminus, an active site sequence with conserved glutamine, valine and glycine residues, and a proline–tryptophan sequence.

Each family of the cystatin superfamily consists of multiple members; cystatin family 2 contains the most members, including cystatins C, S, SN, SA [1] and, more recently, cystatins D, E/M and F [7–10]. The tissue expression of the cystatins varies: some cystatins such as cystatin D exhibit a more restricted tissue distribution [7], whereas others such as cystatin C are more broadly expressed [11]. Although it is well established by studies *in vitro* that the cystatins are cysteine protease inhibitors with specificities against papain-like cysteine proteases such as cathepsins B, S, H and L, the functions of these proteins *in vivo* are not well understood. However, roles in tumour invasion, inflammation and prohormone processing have been proposed [12–14]. Furthermore, several neurological diseases have been associated with mutations in specific cystatin genes, including familial epilepsy as a result of a mutated cystatin B gene [15] and amyloid angiopathy as a result of a single point mutation in the cystatin C gene [16].

Several cystatin-like proteins have been identified that exhibit gene structure, amino acid sequences and/or expression patterns distinct from those of the classic cystatin proteins. These proteins include fetuin [17], cystatin-related proteins [18] and, most recently, testatin [19]. The identification of these proteins has led to the premise that new cystatin families or subgroups of families exist that have evolved to perform tissue-specific functions distinct from the housekeeping-type functions of the classic cystatin proteins.

In previous studies by our laboratory the cystatin-related epididymal spermatogenic (*Cres*) gene was identified and shown to exhibit some sequence similarity with the cystatin family 2 of cysteine protease inhibitors [20]. Unlike the ubiquitous expression of the cystatin C gene, *Cres* gene and protein expression were restricted to a discrete population of germ cells in the testis and the proximal region of the caput epididymidis, suggesting implicit roles in reproduction [20,21]. Recently, we have also determined that the Cres protein is expressed in the male and female anterior

Abbreviations used: *Cres* gene, cystatin-related epididymal spermatogenic gene; 5′ RACE, rapid amplification of 5′ cDNA ends; RT–PCR, reverse transcriptase-mediated PCR.

[1] To whom correspondence should be addressed (e-mail Gail.Cornwall@ttmc.ttuhsc.edu).

The nucleotide sequence data reported will appear in DDBJ, EMBL and GenBank Nucleotide Sequence Databases under the accession numbers AF091503, AF090691, AF090692 and AF059244.

Figure 1 For legend see facing page.

pituitary gonadotropes (G. Sutton and G. A. Cornwall, unpublished work). In contrast with cystatin C, *Cres* expression is highly regulated in these tissues and requires the presence of testicular factors and hormones for expression ([20], and G. Sutton and G. A. Cornwall, unpublished work). The distinctive expression and regulation of the *Cres* gene, taken together with the studies presented here, suggest that the *Cres* gene represents a new subgroup within the family 2 cystatins.

AGAAAAGTAGATGAGAGAGAGAGAAGGACTTTCATATTTCCCAAAGCTTTTAGGAGTTAATCTAAAACATTTTTAATCCAAGTTTCCCCCATTGTTTATTGACAG**ATCACAGACCGAATG** 5777 Exon 3

I T D R M 114 bp

GAATACCAAATTGATGTTCAGATCTCCCGCAGTAATTGCAAAAAACCTTTAAACAATACTGAAAACTGCATCCCTCAAAAAAAACCCGAACTGGAAAAAGTAAGTGACTTCAAGTCCCAT 5897

E Y Q I D V Q I S R S N C K K P L N N T E N C I P Q K K P E L E K

CTAGGGTTATCAAAGCTTCAAGGTAGTTGGGAAATCTTAGGGCGCCCAGCCTAAGATCTCCAAGGGTTCTGGATTGCTCCCCGCTTGGAGTCCCAGAAGGTCCACCTTGCCATCTTTAGA 6017 Intron 3

GTAGAAGATGCAGCAGCATCAAGGTCACTCCTAACCCCAAAACTCCCATCAACTCTGGACTTTTAATGCAGAAAAAACTATTGAAATTTGTTTTCTGAAAGTCTGTCCCCTTGGACTGAG 6137 716 bp

ACTAATTTGTCCAGCTAATGAGGTGAATCATTCATGCTGTGCAGTCATGAAGGAAATCTGCGTGGACAATCATCTTAACTTAGAGCACAGGTACTAGTCATCTGCAGCAGGCCTGTTCCA 6257

AATACAGGAGATTAGACAATATCTAGACCAGATAGCATCCCAAGCTTTCTCTGCACACTCAAGTTTTAAAAGGCTTATCACATCCAGAGAGTAGTAGGTACTGGAGCAAATATGTCAGCC 6377

CTGAGCACGCCCTGACGCCATTTCACATACCTGCTTTGTCATCTACTGTTGTTTGCAGATTGTGATTTTACTTTTATATTTTCTGCTGGGATCCTTGGAATTTTAGGACAGTTGGTGGCA 6497

TGTCTTCTAGCATATTCTTAGGCTAATGCCTAAGAACAGGGCAGTTCTTTTCCTTTTGACATGCTGCTGATGCTCTTTGCTGGTGTCTTTTTCAG**AAAATGAGTTGCAGCTTTCTGGTAG** 6617 Exon 4

K M S C S F L V G 168 bp

GAGCACTTCCCTGGAATGGCGAGTTCAACCTGTTGAGCAAAGAATGTAAAGATGTCTAAGGGTGTCTTGGAGCACCCCTCCCACTTCTGGGTCTGCTTTTCTCTGTAATAAAGAGCAACC 6737

A L P W N G E F N L L S K E C K D V *

TGTGGTTCAAGTCTGTCACCAAC 6857

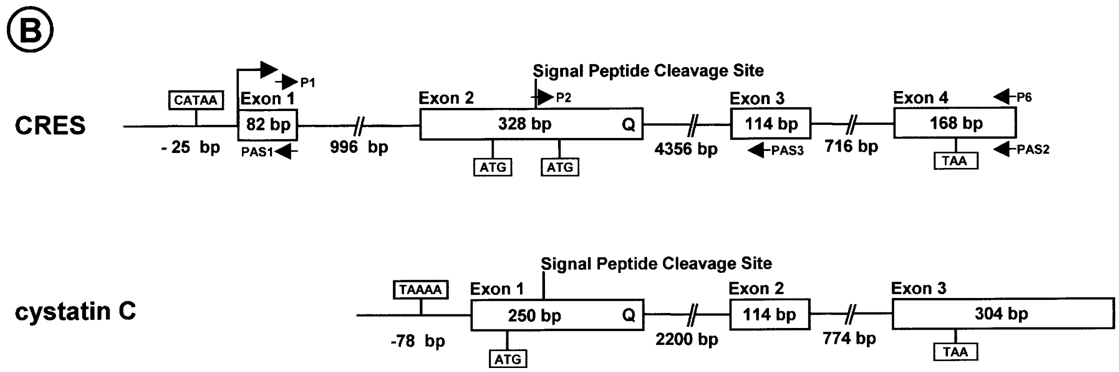

Figure 1 Structural analysis of the mouse *Cres* gene

(**A**) Nucleotide sequence of the mouse *Cres* gene. Exon sequences are in bold type; intron sequences are in normal type. The deduced amino acid sequence is shown below the coding region of the exons. Numbering of nucleotides is based on the major transcription start site as +1 (large arrowhead); a weaker transcription start site at −15 bp is indicated by the small arrowhead. The full-length *Cres* cDNA starts at the major transcription start site. The initiator methionine codon is in exon 2; the first 19 residues represent a putative signal peptide with the signal peptide cleavage site indicated by the vertical arrow. The TATA-like element (CATAA) is boxed. The polyadenylation signal in exon 4 is underlined and a $(TG)_n$ repeat sequence in intron 1 is indicated by a dashed underline. (**B**) Schematic diagram of the *Cres* gene structure compared with the mouse cystatin C gene [2]. The *Cres* gene consists of four exons and three introns as indicated by the boxes and solid lines respectively. Exon 1 and part of exon 2 are a 5′ untranslated sequence with several in-frame stop codons preceding the initiator methionine codon in optimal consensus in exon 2. A second methionine in good consensus is also present in exon 2 and a stop codon and polyadenylation signal are present in exon 4. Q, glutamine. The broken arrow indicates the major transcription start site. P1, P2, P6, oligonucleotide primers used for RT–PCR analysis of *Cres* mRNA. PAS1, PAS2, PAS3, oligonucleotide primers used in 5′ RACE and primer extension studies.

MATERIALS AND METHODS

Animals and reagents

Mature male ICR strain mice were obtained from Harlan (Indianapolis, IN, U.S.A.). Mice were housed under a constant 12 h light/12 h dark cycle and were allowed free access to food and water. All animal studies were performed in accordance with the principles and procedures outined in the NIH Guidelines for Care and Use of Experimental Animals.

Oligonucleotides were commercially prepared by Gibco BRL (Gaithersburg, MD, U.S.A.). Radioisotopes were from New England Nuclear (Boston, MA, U.S.A.) and restriction endonucleases and DNA-modifying enzymes were from Gibco BRL.

Library screening

A C3HF/R1 male mouse lambda 2001 genomic library (kindly provided by C. Wright, Vanderbilt University) was screened for clones with the *Cres* gene by plaque hybridization with a radiolabelled mouse *Cres* cDNA [20]. The *Cres* cDNA was randomly primed with the use of [^{32}P]dCTP and the Prime-It II kit (Stratagene, La Jolla, CA, U.S.A.) and 8×10^5 c.p.m. of probe per ml of hybridization buffer [50 % formaldehyde/5 × SSC/0.2 mg/ml salmon sperm DNA/0.4 mg/ml yeast RNA/50 μg/ml BSA/0.1 % SDS/12.5 mM sodium phosphate buffer (pH 6.6)] were incubated with the filters at 42 °C for 2 days. Filters were washed in 2 × SSC/1 % (w/v) SDS at 42 °C and then at 65 °C before exposure to film. Phage DNA was isolated from the positive clones, subcloned into the Bluescript SK− vector (Stratagene) and sequenced.

Full-length mouse and rat *Cres* cDNA species were obtained by screening mouse testis and rat epididymal cDNA libraries in the UNIZAP vector (Stratagene) by using the original mouse *Cres* cDNA [20] and the hybridization and washing conditions described previously for genomic screening. Positive clones were excised *in vivo* and the resulting phagemid was plated with *Escherichia coli* XL1Bl to prepare plasmid DNA for subsequent sequence analysis. The full-length human *Cres* cDNA was obtained by screening a human testis cDNA library in the ZAP Express vector (Stratagene) under low-stringency hybridization

conditions and the ^{32}P-labelled mouse *Cres* cDNA. In brief, filters were hybridized at 50 °C for 2 days in the presence of 5 × SSC/0.1 % sodium pyrophosphate/50 mM sodium phosphate buffer (pH 6.6)/5 × Denhardts/0.2 mg/ml salmon sperm DNA. Filters were washed in 2 × SSC/0.1 % SDS at 50 °C and then at 65 °C before exposure to film.

Sequence analysis

Double-stranded sequence analysis was performed with α-^{35}S-dATP and the Sequenase 2 sequencing kit (Amersham, Cleveland, OH, U.S.A.). Automated sequencing was performed through the Texas Tech University Biotechnology Core facility under the direction of Dr. Susan San Francisco.

Genomic PCR and genomic walking

To obtain additional *Cres* genomic sequence, PCR was performed with the Elongase system (Gibco BRL) and mouse C3H genomic DNA obtained from the Jackson Laboratories (Bar Harbor, ME, U.S.A.). To obtain the remainder of intron 3 sequences not present in the original genomic clone, a forward primer was generated from intron 3 (5′-AGCACGCCCTGACGCCATTTC-3′) and a reverse primer was generated from *Cres* cDNA (5′-GTGACAGACTTGAACCACAGGTT-3′) sequences. PCR was performed with 500 ng of DNA in the presence of each primer at 200 nM, each dNTP at 200 μM, 60 mM Tris/sulphate, pH 9.1, 18 mM $(NH_4)_2SO_4$, 1.6 mM $MgSO_4$ and 1 μl of Elongase enzyme mix. After an initial denaturation at 94 °C for 30 s, the PCR reaction was incubated at 94 °C for 30 s, 65 °C for 30 s and 68 °C for 7 min (35 cycles) followed by a single 68 °C 15 min incubation with a minicycler (MJ Research, Watertown, MA, U.S.A.). Genomic PCR with the Elongase system and *Cres* cDNA primers was also used to obtain intron 1 (forward, 5′-ACGAGGAAA-CCCAGAAGAACC-3′; reverse, 5′-TGGCATGGAGTATCT-TGTCC-3′) and intron 2 (forward, 5′-CTTGCTATGCACAA-ACCCACTGAGACA-3′; reverse, 5′-ACTGCGGGAGATCT-GAACATC-3′) sequences. PCR products were subcloned into the pGEM-T vector (Promega, Madison, WI, U.S.A.) and sequenced with SP6 and T7 primers.

To obtain *Cres* promoter sequence, the genomic walking method (Clontech, Palo Alto, CA, U.S.A.) was used. Nested anti-sense gene-specific primers generated from *Cres* exon 1 (5′-TTCTGTGACTCTCCCACCTGCTCGT-3′ and 5′-TGGCCA-CCCCAGCCTTGTGGTTCTT-3′) were used with adaptor primers supplied by the manufacturer and Advantage Tth polymerase mix (Clontech) to hot-start PCR amplify mouse genomic fragments with ligated adaptor sequences. PCR was performed with an initial denaturation at 94 °C for 1 min to inactivate the TthStart antibody followed by 94 °C for 2 s, 72 °C for 3 min (6 cycles) and then 94 °C for 2 s, 67 °C for 3 min (31 cycles) followed by 67 °C for 4 min. The resulting PCR products were subcloned into the pGEM-T vector and sequenced with SP6 and T7 primers.

Primer extension and rapid amplification of 5′ cDNA ends (5′ RACE)

For primer extension studies, 30 μg of total RNA from the mouse proximal caput epididymidis were incubated with a ^{32}P end-labelled *Cres* anti-sense primer (PAS1) (5′-TTCTGTGAC-TCTCCCACCTGCTCGT-3′) in the presence of 10 mM Tris/HCl, pH 8.3, 90 mM KCl, 1 mM $MnCl_2$, each dNTP at 200 μM and 0.25 unit/μl rTh reverse transcriptase (Perkin–Elmer, Foster City, CA, U.S.A.) at 70 °C for 1 h. cDNA products were analysed on an 8 % (w/v) polyacrylamide/6 M urea gel in TBE buffer. Control reactions were performed in the absence of RNA. End-labelling was performed by incubation of 10 pmol *Cres* primer in the presence of 50 mM Tris/HCl, pH 7.6, 10 mM $MgCl_2$, 5 mM dithiothreitol, 100 μM spermidine, 30 μCi of [γ-^{32}P]dATP and 10 units of T4 polynucleotide kinase at 37 °C for 10 min, followed by 90 °C for 2 min to inactivate the enzyme. End-labelled products were purified by centrifugation through a G-25 Sephadex spin column (Boehringer Mannheim, Indianapolis, IN, U.S.A.) to remove free nucleotide.

5′ RACE was performed with the 5′ RACE system (Gibco BRL). Total RNA (1 μg) from the mouse proximal caput epididymidis was reverse-transcribed with a *Cres* anti-sense primer (PAS2) (5′-GTGACAGACTTGAACCACAGGTT-3′) and Superscript II reverse transcriptase (Gibco BRL) for 50 min at 42 °C. Reactions were incubated with RNAse mix (RNAse H and RNAse T1) for 30 min at 37 °C and then centrifuged through Glassmax spin columns. Purified cDNA species were dC-tailed and PCR amplified with the abridged anchor primer supplied by the manufacturer and a second nested *Cres* gene-specific primer (PAS3) (5′-ACTGCGGGAGATCTGAACATC-3′) in the presence of *Taq* DNA polymerase (Sigma Chemical Co., St. Louis, MO, U.S.A.). PCR was performed at 94 °C for 30 s, 55 °C for 30 s, 72 °C for 1 min (30 cycles). 5′ RACE products were analysed by 1 % (w/v) agarose-gel electrophoresis in the presence of TAE buffer (1 × TAE is 0.04 M Tris/0.001 M EDTA/0.11 % acetic acid) and then subcloned into the pGEM-T Easy vector (Promega) and sequenced.

Preparation of RNA

Total RNA was isolated from mouse tissues and LβT2 gonadotrope cells (a gift from P. Mellon) with the use of Trizol reagent (Gibco BRL) in accordance with the manufacturer's protocol.

Reverse transcriptase-mediated PCR (RT–PCR)

RT–PCR was performed with 1 μg of total RNA from mouse testis, epididymis and LβT2 cells, and 2 μg of RNA from mouse pituitary gland. RNA was incubated in reverse transcription reaction buffer containing 5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris/HCl, pH 8.3, 0.5 mM dNTPs, 20 units of RNAsin and 2.5 μM oligo(dT) for 30 min at 37 °C in the presence of 3 units of DNAse I (Boehringer Mannheim). After heat inactivation of the DNAse I at 75 °C for 5 min, 50 units of MuLV reverse transcriptase (Perkin–Elmer) was added and the reaction mixture was incubated at 42 °C for 15 min, followed by 99 °C for 5 min to inactivate the enzyme. An 18 μl portion of the 20 μl reverse transcription reaction was used for the PCR amplification of the *Cres* cDNA in the presence of 2 mM $MgCl_2$, 10 mM Tris/HCl, pH 8.3, 50 mM KCl, 0.2 mM dNTPs, 0.5 μM forward and reverse *Cres*-specific primers and 1.25 units of *Taq* DNA polymerase (Sigma Chemical Co.). After an initial denaturation at 95 °C, PCR was performed at 94 °C for 1 min, 65 °C for 1 min (36 cycles) followed by 72 °C for 7 min with a minicycler (MJ Research). RT–PCR products were analysed by agarose-gel electrophoresis followed by staining with ethidium bromide. Oligonucleotide primers for PCR were designed from the mouse *Cres* cDNA: P1 (forward), 5′-GCAGGTGGAGAGTCACAG-AA-3′; P2 (forward), 5′-CTGGCAGTTGGTGTGGATCA-3′; P6 (reverse), 5′-GTGACAGACTTGAACCACAGGTT-3′.

Chromosomal mapping

Chromosomal mapping was performed with the assistance of the Jackson Laboratory (Bar Harbor, ME, U.S.A.). To identify

restriction enzymes that would produce restriction fragment length polymorphisms, genomic DNA from C57BL6/J and *M. spretus* mice was digested with restriction enzymes *Bam*HI, *Bgl*II, *Eco*RV, *Hin*dIII, *Msp*I, *Pst*I, *Pvu*II, *Sst*I, *Taq*I and *Xba*I followed by Southern blot analysis. A 200 bp 5′ *Cres* cDNA probe was prepared by random-primer labelling as described above. Mapping was performed by restriction digestion with *Msp*I and Southern blot analysis on the 94 mice from the Jackson Laboratory interspecies backcross panel (C57BL6/JEi × SPRET/Ei)F_1 × SPRET/Ei mice, known as the Jackson BSS panel [22].

RESULTS AND DISCUSSION

Isolation and structural analysis of the *Cres* gene

Screening of a mouse genomic library with a random-primed *Cres* cDNA [20] identified clones containing *Cres* sequences. A sequence analysis of one clone showed that the genomic DNA fragment contained the complete exon 3 sequence flanked 5′ and 3′ by intron sequences. Two PCR strategies were then utilized to obtain the remainder of the *Cres* gene sequence and the *Cres* promoter and 5′ flanking sequences. These studies show that the *Cres* gene is approx. 6.8 kb in size. A comparison of the genomic sequence with the mouse cDNA sequence revealed the presence of four exons and three introns (Figure 1). A computer-generated analysis (Splice Site Prediction by Neural Network) of the *Cres* exon/intron splice junctions showed these sites to be common donor/acceptor sequences observed in other genes. Exon 1 and part of exon 2 consist of 5′ untranslated sequence with an initiator methionine codon in optimal context and a downstream methionine codon in good context [23] present in exon 2 (Figure 1B). On the basis of translational initiation at the first ATG there is a 19-residue signal peptide sequence with the cleavage site predicted to be between Ala^{19} and Val^{20} following established rules for predicting signal sequence cleavage sites [24]. A stop codon and polyadenylation signal are present in exon 4.

A comparison of the *Cres* gene structure with the structure of family 2 cystatin genes revealed several similarities (Figure 1B). Although the *Cres* gene has an extra exon (exon 1), the remaining exons of the *Cres* gene containing the coding sequence are similar in size to that of the mouse cystatin C gene [2]. Furthermore, the exon/intron splicing occurs in the same location within the coding sequence of the two genes. For example, in the cystatin C gene, a splice site occurs within one of the consensus sites (Q/-V-G) in cystatin exon 1, which is thought to be critical for binding to cysteine proteases. Although the *Cres* sequence in this region is distinct, the splice site is identical with that of the cystatin C gene and occurs immediately after the glutamine residue in *Cres* exon 2 (Figure 1B). The possibility that splicing might occur before the glutamine residue cannot, however, be excluded because the same nucleotides (CAG) are at both the 5′ and 3′ termini of *Cres* intron 2 and the resulting mRNA would not differ if either splicing event occurred. Splicing before the glutamine residue, although possible, is unlikely because the splice junction would then be unconventional. The splicing of *Cres* exon 3 occurs 10 residues after a conserved cysteine residue and this splice site is in the same location within the coding region of the cystatin C gene, cystatin C exon 2 [2]. The sizes of the intron sequences, however, are quite different between the *Cres* and cystatin C genes. Unlike the relatively small size of the cystatin C introns [2], the *Cres* gene has a large intron 2 with smaller introns 1 and 3. Another distinctive feature of the *Cres* gene is the presence of a $(TG)_{14–20}$ stretch in intron 1 (Figure 1A). This element is also present in the introns of the cystatin C gene [2] and in a cystatin-related protein gene [25] as well as in the promoters and/or introns of other genes not related to the cystatins, including the glutathione peroxidase-like [26] and CRISP-1 genes [27]. These elements have the potential to adopt alternative DNA conformations that might direct the binding of transcription factors or other regulatory proteins to sites in the promoter [28].

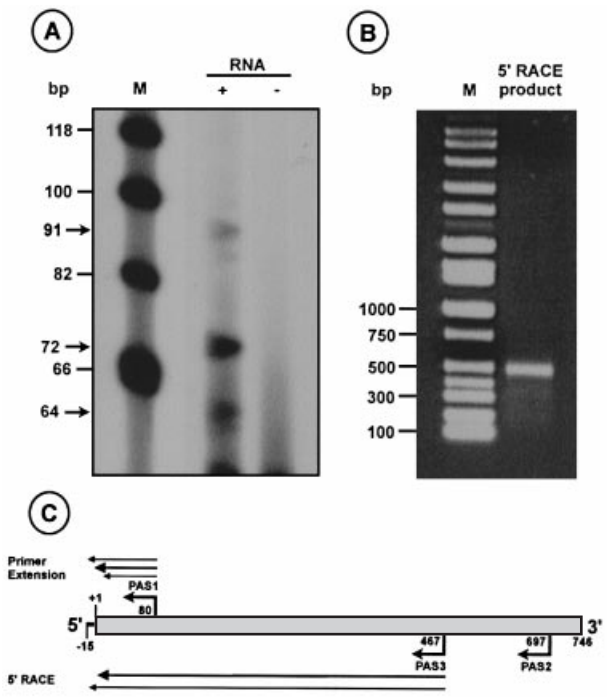

Figure 2 Determination of *Cres* transcription start sites by primer extension and 5′ RACE assays

(**A**) Primer extension was performed with total RNA from the mouse proximal caput epididymidis and a ^{32}P end-labelled *Cres* anti-sense primer (PAS1) designed from the 5′ region of the *Cres* cDNA (+). Control reactions were performed in the absence of RNA (−). The extended products were separated on an 8% (w/v) polyacrylamide/6 M urea gel followed by exposure to film. Three primary extended products of 91, 72 and 64 bp were identified, with the 72 bp product being predominant. (**B**) 5′ RACE was performed with total RNA from the mouse proximal caput epididymidis. RNA was reverse transcribed and the cDNA species were PCR amplified with two nested *Cres* anti-sense primers (PAS2, PAS3) followed by analysis of the 5′ RACE products by agarose-gel electrophoresis and staining with ethidium bromide. A predominant PCR product of approx. 470 bp was observed. The PCR products were subcloned into the pGEM-T Easy vector and twelve clones were sequenced. (**C**) Schematic diagram of the *Cres* cDNA showing the location of the *Cres* anti-sense primer (PAS1) and extended products (horizontal arrows with thicker arrow indicating the predominant product) resulting from primer extension assays (top) and the location of the *Cres* anti-sense primers (PAS2, PAS3) and 5′ RACE products (horizontal arrows with thicker arrow indicating the predominant product) generated from 5′ RACE (bottom). The predominant primer extended product (72 bp) and most of the 5′ RACE products placed transcriptional initiation at the adenine residue at the beginning of the full-length *Cres* cDNA. A weaker transcriptional start site was also identified 15 bp upstream of the major start site. The autoradiograms and photographs of the ethidium bromide stained gels were scanned with a Mirror 1200 colour scanner and the results were transferred into Adobe Photoshop 4.0.

Analysis of the *Cres* promoter

An examination of the *Cres* promoter revealed the presence of a TATA-like element (CATAA) approx. 25 bp upstream of the major predicted transcription start site (Figure 1A), which was confirmed by primer extension and 5′ RACE assays. As shown in Figure 2(A), primer extension with total RNA isolated from the mouse proximal caput epididymidis and an anti-sense primer (PAS1) designed from the 5′ end of the *Cres* cDNA resulted in three extended products. The most abundant 72 bp extended

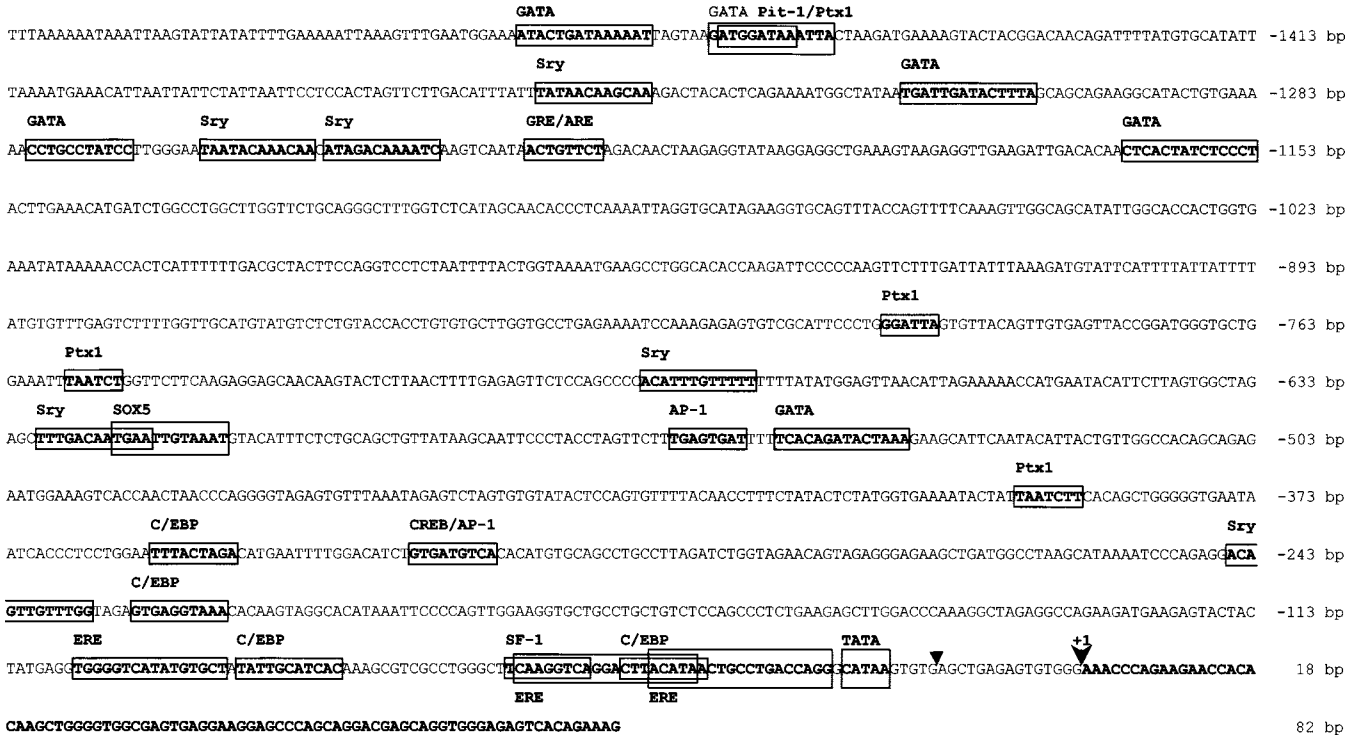

Figure 3 Analysis of the *Cres* promoter and 5′ flanking sequences

The 1527 bp *Cres* 5′ flanking sequences were examined for DNA-binding protein consensus sites with the computer program MATINSPECTOR [30]. The TATA-like element (CATAA) is indicated and the major transcription start site of exon 1 is shown by the large arrowhead (+1); the weaker transcription start site is indicated by the small arrowhead. The consensus motifs for *cis*-elements characterized in other gene promoter and enhancer sequences are boxed. Abbreviations: GATA, GATA-binding factor 1; Ptx1, pituitary homeobox 1; Sry, sex-determining region Y gene product; GRE, glucocorticoid response element; ARE, androgen response element; Sox5, Sry-related high mobility group (HMG) box; AP-1, activator protein 1; C/EBP, CCAAT/enhancer-binding protein; CREB, cAMP-response-element-binding protein; ERE, oestrogen response element; SF-1, steroidogenic factor 1.

product predicted the major transcription start site to be at or near the beginning of the mouse full-length cDNA (Figure 2C; Figure 1A, large arrowhead). The longer and less abundant product of 91 bp predicts a weaker transcription start approx. 15 bp further upstream (Figure 2C; Figure 1A, small arrowhead). The shorter, less abundant product (64 bp) probably represents premature stalling of the reverse transcriptase enzyme. The transcription start sites predicted by primer extension assays were confirmed by 5′ RACE. 5′ RACE with nested *Cres* antisense primers (PAS2 and PAS3) (Figure 2C) resulted in a predominant PCR product of the expected size on the basis that the transcription start was at the beginning of the full-length *Cres* cDNA (Figure 2B). Sequence analysis of the cloned PCR products showed that most of the PCR products terminated at the major transcription start site predicted by the primer extension assays. This placed transcriptional initiation at an adenine residue, which is typical of most eukaryotic genes [29] (Figure 1A, large arrowhead). One of the PCR products, however, represented an mRNA initiated from an adenine residue at the weaker transcription start site approx. 15 bp upstream of the major start site.

An examination of the *Cres* promoter and 5′ flanking sequences with a computer program to predict consensus DNA/protein-binding sites [30] revealed the presence of putative binding sites for transcription factors known to interact with the promoters of regulated genes (Figure 3). There were several DNA motifs characteristic of genes involved in steroidogenesis and reproductive function, including three oestrogen response elements

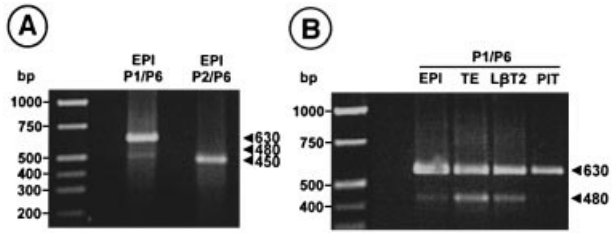

Figure 4 Alternative splicing of the *Cres* mRNA

(**A**) Ethidium bromide-stained agarose gel of the RT–PCR products generated from mouse caput epididymal RNA (EPI) with the P1/P6 and P2/P6 primers. RT–PCR with the P1/P6 primers resulted in two PCR products of 630 and 480 bp, whereas RT–PCR with the P2/P6 primers resulted in a single PCR product of 450 bp. (**B**) Ethidium bromide-stained agarose gel of RT–PCR products generated from mouse caput epididymis (EPI), testis (TE), LβT2 gonadotrope cell line and pituitary (PIT) RNA with the P1/P6 primers. In all tissues the predominant RT–PCR product was the larger 630 bp product. Photographs of the ethidium bromide-stained gels were scanned with a Mirror 1200 colour scanner and the data were transferred into Adobe Photoshop 4.0.

[31] near the *Cres* transcription start site (−41, −59 and −106 bp) as well as an androgen response element/glucocorticoid response element [32] half site at −1229 bp. A single steroidogenic factor 1 [33] site with 100% conservation was also present at −60 bp as well as a Pit-1 [34] site at

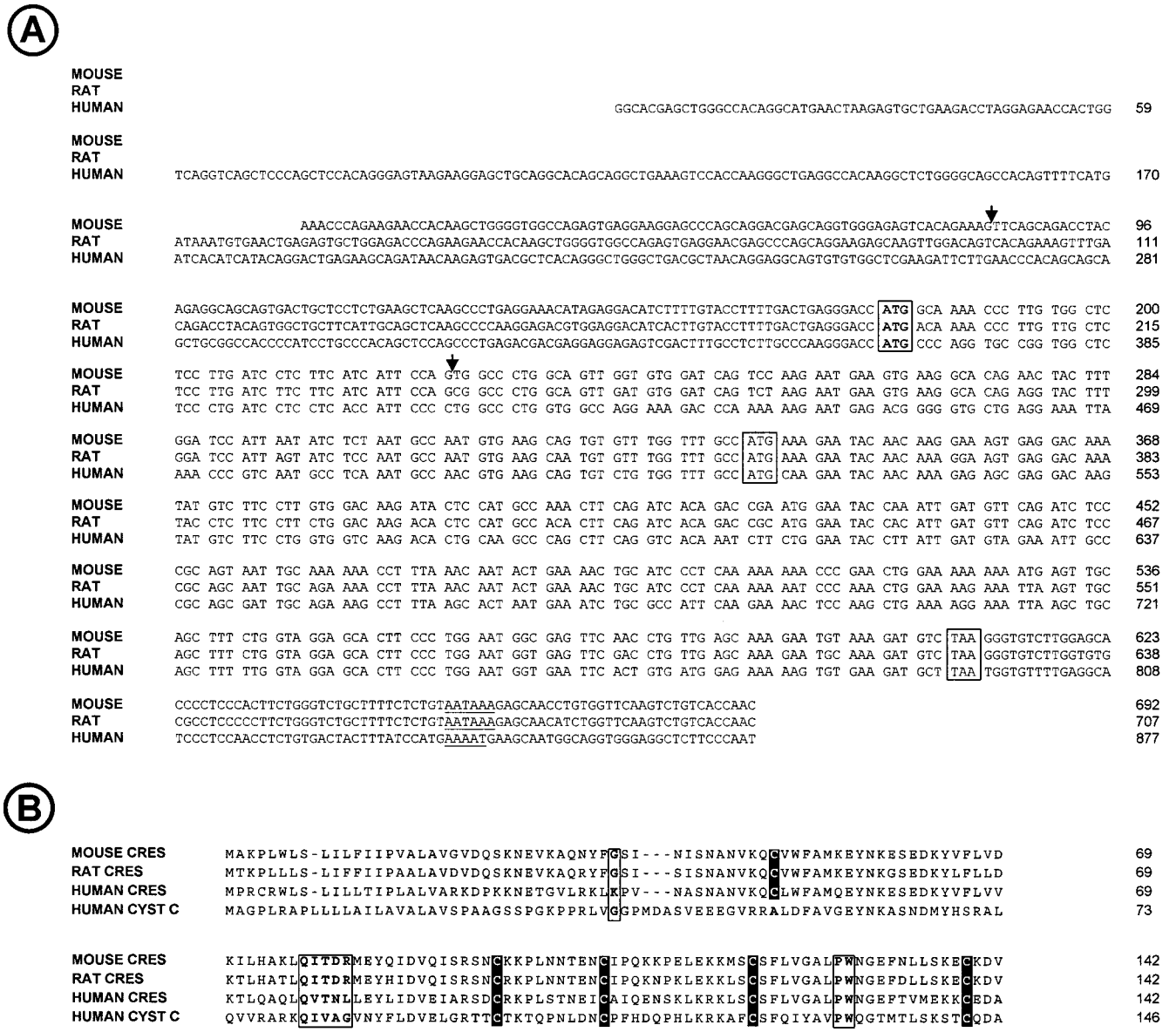

Figure 5 Comparison of *Cres* and cystatin C cDNA and amino acid sequences

(**A**) Nucleotide sequence of the full-length mouse, rat and human *Cres* cDNA species. The 5′ untranslated, full coding and 3′ untranslated sequences are shown. The initiator and downstream methionine codon and the stop codon are boxed. The polyadenylation signal is underlined. The sequence between the arrows is missing from the mouse splice variant cDNA. The overall nucleotide sequence similarity in the coding region between the mouse and rat *Cres* cDNA species is 91% and between mouse and human is 73%. (**B**) Derived amino acid sequences for the mouse, rat and human Cres proteins compared with the human cystatin C protein. The three consensus motifs considered necessary for cystatin function are boxed. Cysteine residues are in white on a black background. The derived amino acid sequence similarity between mouse and rat *Cres* protein is 83% and between mouse and human Cres protein is 58%. The sequence similarity between human Cres and cystatin C proteins is 27%. Broken lines represent gaps introduced for alignment.

−1469 bp and four pituitary homeobox 1 [35] sites at positions −398, −756, −802 and −1469. In addition, seven consensus sites that bind the sex-determining region Y transcription factor [36] and its related DNA-binding protein Sox-5 [37] were present throughout the 1.6 kb of *Cres* promoter sequences. Other DNA-binding sites in the *Cres* promoter included those for CCAAT/enhancer-binding protein [38], cAMP-response-element-binding protein [39] and GATA-1-binding protein [40]. Considering the highly restricted expression of the *Cres* gene to the testicular haploid germ cells, epididymis [20] and the anterior pituitary gonadotropes (G. Sutton and G. A. Cornwall, unpublished work), the presence of some of these DNA-binding elements is not surprising and might suggest a means by which the *Cres* gene is regulated. However, whether these sites are functional remains to be determined. Studies are currently in progress to assess these DNA elements in *Cres* gene regulation.

Alternative splicing of the *Cres* mRNA

During the course of using RT–PCR to examine *Cres* gene expression in mouse tissues, it was observed that one set of primer pairs (P1, P6) designed from the original *Cres* cDNA

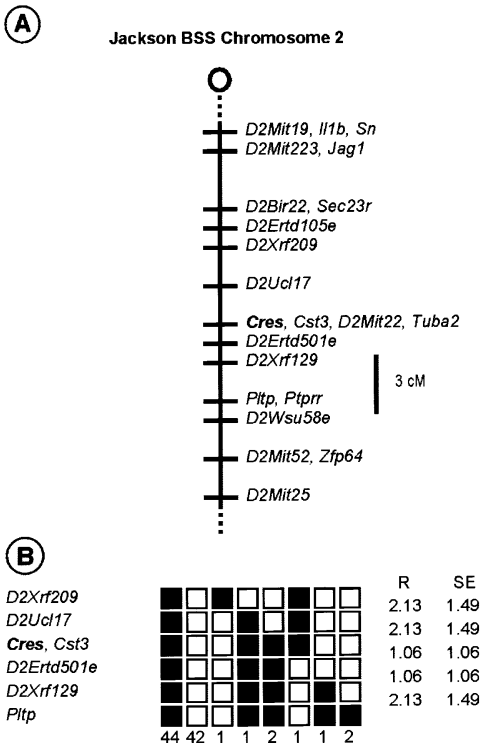

Figure 6 Chromosomal mapping of the mouse *Cres* gene

(**A**) Map figure from the Jackson BSS backcross showing part of mouse chromosome 2. The map is depicted with the centromere towards the top. A 3 cM scale bar is shown at the right. Loci mapped to the same position are listed in alphabetical order. Missing typings were inferred from surrounding data where assignment was unambiguous. Raw data from the Jackson Laboratory were obtained from the WWW address http://www.jax.org/resources/documents/cmdata. Abbreviations: *Cres*, cystatin-related epididymal spermatogenic; *Cst 3*, cystatin C [2]. (**B**) Haplotype figure from the Jackson BSS backcross showing part of chromosome 2 with loci linked to *Cres*. Loci are listed in order, with the most proximal at the top. The black boxes represent the C57BL6/JEi allele and the white boxes the SPRET/Ei allele. The number of animals with each haplotype is given at the bottom of each column of boxes. The percentage recombination (*R*) between adjacent loci is given to the right of the Figure with the S.E.M. (SE) for each *R*. Missing typings were inferred from surrounding data where assignment was unambiguous.

resulted in two PCR products, whereas another primer pair (P2, P6) resulted in a single PCR product from the same mouse caput epididymal RNA sample (Figure 4A). This was shown not to represent a priming artifact of the primers from the mRNA because PCR of the cDNA with the (P1, P6) primer pair resulted in a single PCR product of the expected size (results not shown). To determine whether the PCR products represented different *Cres* mRNA species, the PCR products were cloned and sequenced. The larger PCR product contained an additional 143 bp sequence not present in the smaller PCR product. A comparison of the PCR sequences with the *Cres* genomic sequence revealed that the 5′ half of exon 2 was present neither in the smaller PCR product nor in the original cDNA sequence (Figure 1B). Additional library screening of mouse testis, rat epididymal and human testis cDNA libraries identified *Cres* cDNA species of 692, 707 and 877 bp respectively, all of which when sequenced contained the 143 bp 5′ half of exon 2 and therefore seemed to represent the full-length and correctly spliced *Cres* mRNA species (Figures 5A and 5B). An examination of the sequence surrounding the alternative splice site in the middle of exon 2 showed a consensus intron/exon splice junction with a predicted splicing score (0.78) slightly less than the splicing score for the correct intron 1/exon 2 splice junction (0.87). Therefore, because the 5′ region of exon 2 is a conserved intron/exon splice sequence, both splice sites are utilized in the processing of the *Cres* mRNA and two *Cres* mRNAs are produced. However, an analysis of the relative proportions of the two *Cres* mRNA species in tissues that express the *Cres* gene revealed that most of the *Cres* mRNA was spliced at the first splice site and thus the splice variant mRNA was a minor form (Figure 4B).

The functional significance of the alternative splicing of the *Cres* mRNA has yet to be determined. Because the splice variant mRNA lacks the 5′ half of exon 2 it lacks the initiator methionine codon and therefore if translated would utilize a downstream ATG codon in exon 2. The resulting truncated protein would be approx. 12 kDa and would lack the signal sequence as well as additional amino acids present in the mature Cres protein. Preliminary transcription and translation studies *in vitro* indicate that translational initiation can occur at the second ATG in the *Cres* sequence (results not shown). Therefore the utilization of different ATG codons might be a means of regulating Cres protein function. Interestingly, alternative splicing of the cathepsin B mRNA also results in deletion of the signal peptide sequence with subsequently a unique intracellular localization of the protein [41].

Chromosomal mapping of the *Cres* gene

To determine the chromosomal localization of the *Cres* gene, genetic mapping studies were performed. Southern blot analysis of DNA from 94 backcrosses derived from the matings of $(C57BL6/Ei \times SPRET/Ei)F_1 \times SPRET/Ei$ mice was performed by using the mouse *Cres* cDNA. The allele pattern of *Cres* was compared with approx. 4000 other loci previously mapped in the Jackson BSS cross and linkage was found to the distal region of chromosome 2 (Figures 6A and 6B). The Mouse Nomenclature Committee of the Jackson Laboratory (Bar Harbor, ME, U.S.A.) has approved this locus *Cres* (MGD accession number J:48910). The mapping studies also showed that the *Cres* gene co-segregated with the cystatin C (*Cst 3*) locus. The mapping of the *Cres* and cystatin C genes to the same genetic loci is not surprising because the *Cres* gene seems to be evolutionarily related to the cystatin family 2 members. Several of the cystatin family 2 genes, including cystatins C, D, S, SN and SA [3], have been shown to cluster on human chromosome 20, which exhibits a region of linkage conservation with the distal region of mouse chromosome 2.

Evolutionary relationship of *Cres* to the cystatin superfamily

The similarity of the *Cres* genomic structure and splice junctions with the cystatin C gene as well as the co-segregation of *Cres* with cystatin C on mouse chromosome 2 indicate that the *Cres* gene is most probably derived from the same ancestral gene as the cystatin C gene. A phylogenetic tree constructed with the MegAlign program of the Lasergene Software Suite (DNA STAR, Madison, WI, U.S.A.) formally demonstrates the evolutionary relationship between the mature protein sequences of *Cres* and the cystatin families (Figure 7). The alignment suggests that Cres protein, although probably a new member of the family 2 cystatins, is distinct from the cystatins C, D and S and might therefore represent a new subgroup within the family. Cres protein, like the family 2 cystatins, contains the proline–tryptophan (PW) residues thought to participate in the interaction of cystatins with cysteine proteases (Figure 5B). Although the Cres sequence also contains the glutamine (Q) residue at the appropriate site of a second consensus motif for

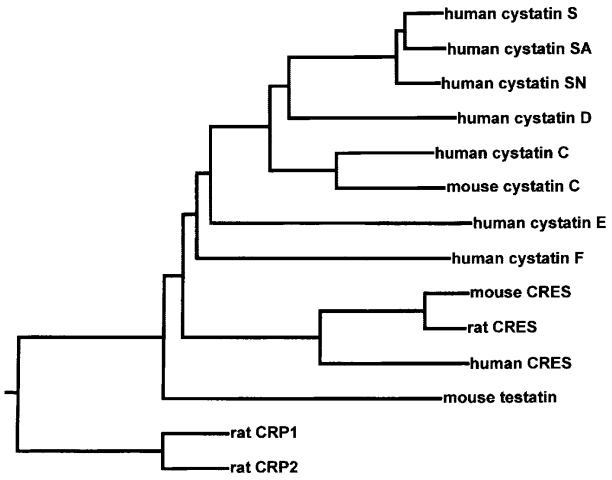

Figure 7 Evolutionary relationship of the Cres protein to the cystatin superfamily

The phylogenetic tree was generated with the MegAlign program of the Lasergene Software Suite (DNA STAR). Signal sequences were removed before alignment. Only the cystatin domains in the kininogens were used in the alignment. The construction of the tree utilized the clustal method with the PAM 250 alignment matrix. The multiple alignment parameters were: gap penalty, 20; gap length penalty, 20. The pairwise alignment parameters were: ktuple, 1; gap penalty, 3.

inhibition (Q-V-G), the other residues are quite unique from that in the cystatin C sequence. Finally, whereas mouse and rat Cres sequences possess a glycine residue in an appropriate position at the N-terminus to form a third consensus motif for cystatin function, the human Cres sequence does not (Figure 5B). Cres protein is also like the mammalian single domain cystatin proteins in that it contains both the type A and B disulphide loops with the appropriate spacing between the cysteine residues. In all three species the Cres sequence also possesses a third cysteine residue in the N-terminus, suggesting that Cres might contain half of a characteristic type C disulphide loop that is present in insect cystatins as well as in a cystatin domain of the three cystatin domain proteins fetuin and kininogen.

Taken together, these studies suggest that the *Cres* gene represents a new subgroup within the family 2 cystatins. However, unlike the proposed function of the cystatins as general regulators of cysteine protease activities, the highly regulated and tissue-specific expression of the *Cres* gene implies that it might perform very specific functions in the reproductive tract. On the basis of our previous studies that Cres protein is present in the luminal fluid of the proximal caput epididymidis [20] as well as our recent observations that Cres protein co-localizes with LHβ protein to the anterior pituitary gonadotrope secretory granules (G. Sutton and G. A. Cornwall, unpublished work) as well as to the sperm acrosomes, a lysosome-like structure (P. Syntin and G. A. Cornwall, unpublished work), suggests that Cres protein might have roles in the regulation of prohormone and sperm acrosomal and plasma membrane protein processing. Furthermore, our finding that there is an alternatively spliced *Cres* mRNA lacking the signal sequence suggests that Cres protein might function intracellularly as well as extracellularly.

We thank Simon Williams and Steve Hann for their critical reading of the manuscript. These studies were supported by NIH grant HD33903 (G. A. C.).

Received 14 September 1998/25 January 1999; accepted 23 February 1999

REFERENCES

1 Turk, V. and Bode, W. (1991) FEBS Lett. **285**, 213–219
2 Huh, C., Nagle, J. W., Kozak, C. A., Abrahamson, M. and Karlsson, S. (1995) Gene **152**, 221–226
3 Thiesse, M., Miller, S. J. and Dickinson, D. P. (1994) DNA Cell Biol. **13**, 97–116
4 Nycander, M. and Bjork, I. (1990) Biochem. J. **271**, 281–284
5 Bode, W., Engh, R., Musil, D., Thiele, U., Huber, R., Karshikov, A., Brzin, J., Kos, J. and Turk, V. (1988) EMBO J. **7**, 2593–2599
6 Stubbs, M. T., Laber, B., Bode, W., Huber, R., Jerala, R., Lenarcic, B. and Turk, V. (1990) EMBO J. **9**, 1939–1947
7 Freije, J. P., Abrahamson, M., Olafsson, I., Velasco, G., Grubb, A. and Lopez-Otin, C. (1991) J. Biol. Chem. **266**, 20538–20543
8 Ni, J., Abrahamson, M., Zhang, M., Fernandez, M. A., Grubb, A., Su, J., Yu, G.-L., Li, Y., Parmelee, D., Xing, L. et al. (1997) J. Biol. Chem. **272**, 10853–10858
9 Sotiropoulou, G., Anisowicz, A. and Sager, R. (1997) J. Biol. Chem. **272**, 903–910
10 Ni, J., Fernandez, M. A., Danielsson, L., Chillakuru, R. A., Zhang, J., Grubb, A., Su, J., Gentz, R. and Abrahamson, M. (1998) J. Biol. Chem. **273**, 24797–24804
11 Abrahamson, M., Olafsson, I., Palsdottir, A., Ulvsback, M., Lundwall, A., Jensson, O. and Grubb, A. (1990) Biochem. J. **268**, 287–294
12 Sloane, B. F., Rozhin, J., Robinson, D. and Honn, K. V. (1990) Biol. Chem. Hoppe-Seyler **371**, 193–198
13 Leung-Tack, J., Tavera, C., Martinez, J. and Colle, A. (1990) Inflammation **14**, 247–258
14 Orlowski, M. (1983) Mol. Cell. Biochem. **52**, 49–74
15 Pennachio, L. A., Lehesjoki, A.-E., Stone, N. E., Willour, V. L., Virtaneva, K., Maiao, J., D'Amato, E., Ramirez, L., Faham, M., Koskiniemi, M. et al. (1996) Science **271**, 1731–1734
16 Ghiso, J., Jensson, O. and Frangione, B. (1986) Proc. Natl. Acad. Sci. U.S.A. **83**, 2974–2978
17 Kellerman, J., Haupt, H., Auerswald, E.-A. and Muller-Esterl, W. (1989) J. Biol. Chem. **264**, 14121–14128
18 Winderickx, J., Hemschoote, K., De Clercq, N., Van Dijck, P., Peeters, B., Rombauts, W., Verhoeven, G. and Heyns, W. (1990) Mol. Endocrinol. **4**, 657–667
19 Tohonen, V., Osterlund, C. and Nordqvist, K. (1998) Proc. Natl. Acad. Sci. U.S.A. **95**, 14208–14213
20 Cornwall, G. A., Orgebin-Crist, M.-C. and Hann, S. R. (1992) Mol. Endocrinol. **6**, 1653–1664
21 Cornwall, G. A. and Hann, S. R. (1995) Mol. Reprod. Dev. **41**, 37–46
22 Rowe, L. B., Nadeau, J. H., Turner, R., Frankel, W. N., Letts, V. A., Eppig, J. T., Ko, M. S. H., Thurston, S. J. and Birkenmeier, E. H. (1994) Mammalian Genome **5**, 253–274
23 Kozak, M. (1989) J. Cell Biol. **108**, 229–241
24 von Heijne, G. (1986) Nucleic Acids Res. **14**, 4683–4689
25 Devos, A., De Clercq, N., Vercaeren, I., Heyns, W., Rombauts, W. and Peeters, B. (1993) Gene **125**, 159–167
26 Ghyselinck, N. B., Dufaure, I., Lareyre, J.-J., Rigaudiere, N., Mattei, M.-G. and Dufaure, J.-P. (1993) Mol. Endocrinol. **7**, 258–272
27 Schwidetzky, U., Schleuning, W.-D. and Haendler, B. (1997) Biochem. J. **321**, 325–332
28 Hamada, H., Petrino, M. G. and Kakunaga, T. (1982) Proc. Natl. Acad. Sci. U.S.A. **79**, 6455–6469
29 Breathnach, R. and Chambon, P. (1981) Annu. Rev. Biochem. **50**, 349–383
30 Quandt, K., Frech, K., Wingender, E. and Werner, T. (1995) Nucleic Acids Res. **23**, 4878–4884
31 Klein-Hitpass, L., Ryffel, G. U., Heitlinger, E. and Cato, A. C. (1988) Nucleic Acids Res. **16**, 647–663
32 Ham, J., Thomson, A., Needham, M., Webb, P. and Parker, M. (1988) Nucleic Acids Res. **16**, 5263–5276
33 Lala, D. S., Rice, D. A. and Parker, K. L. (1992) Mol. Endocrinol. **6**, 1249–1258
34 Ingraham, H. A., Chen, R., Mangalam, H. J., Elsholtz, H. P., Flynn, S. E., Lin, C. R., Simmons, D. M., Swanson, L. and Rosenfeld, M. G. (1988) Cell **55**, 519–529
35 Lamonerie, T., Trembaly, J. J., Lanctot, C., Therrien, M., Gauthier, Y. and Drouin, J. (1996) Genes Dev. **10**, 1284–1295
36 Haqq, C. M., King, C. Y., Donahoe, P. K. and Weiss, M. A. (1993) Proc. Natl. Acad. Sci. U.S.A. **90**, 1097–1101
37 Denny, P., Swift, S., Connor, F. and Ashworth, A. (1992) EMBO J. **11**, 3705–3712
38 Johnson, P. F. (1993) Mol. Cell. Biol. **13**, 6919–6930
39 Meyer, T. E. and Habener, J. F. (1993) Endocr. Rev. **14**, 269–290
40 Evans, T., Reitman, M. and Felsenfeld, G. (1988) Proc. Natl. Acad. Sci. U.S.A. **85**, 5976–5980
41 Mehtani, S., Gong, Q., Panella, J., Subbiah, S., Peffley, D. M. and Frankfater, A. (1998) J. Biol. Chem. **273**, 13236–13244